基于大语言模型的自杀意念文本数据增强与识别技术研究*

章彦博[1,2] 黄峰[1,2,3] 莫柳铃[4] 刘晓倩[1,2] 朱廷劭[1,2]

（[1] 中国科学院心理研究所行为科学重点实验室，北京 100101）

（[2] 中国科学院大学心理学系，北京 100101）

（[3] 香港城市大学计算学院数据科学系，香港 999077）

（[4] 南开大学社会学院社会心理学系，天津 300350）

摘　要 自杀已成为全球性公共卫生难题，传统的自杀意念识别方法主要依赖患者主动求助，而基于文本分析的自动识别模型则受限于标注数据的稀缺性。本研究创新性地提出一种基于大语言模型的数据增强方法，旨在提升自杀意念文本识别的精度。研究采用双阶段设计：研究一聚焦于数据增强，研究二验证增强效果。在研究一中，选用 ChatGLM3-6B 和 Qwen-7B-Chat 作为底层模型，结合有监督学习策略与零样本和少样本学习方法，优化训练数据集质量。通过 8 组严谨的对比实验，结果显示两类自研模型在数据增强方面表现卓越，其处理后数据集的综合得分分别达到 0.90 和 0.92，显著优于基线模型（$p < 0.001$）。研究二进一步评估了数据增强对识别模型性能的影响，结果表明，增强后的模型在识别准确率和正确拒绝率指标上全面超越最佳基线模型（$p < 0.001$）。本研究不仅验证了基于大语言模型的数据增强方法在提升自杀意念识别模型性能方面的有效性，还为心理健康领域的人工智能应用开辟了新方向。这种方法有望在保护用户隐私的同时，提供及时、有效的自杀风险早期预警，为自杀预防工作提供重要的技术支持和研究思路。未来研究可着眼于扩大数据异质性、优化提示工程设计、引入人机交互范式等，进一步拓展该方法在促进临床心理诊断领域的应用。

关键词 自杀意念，数据增强，自杀文本识别，大语言模型，人工智能

1 引言

自杀作为全球性的公共卫生挑战，对人类社会构成巨大威胁。根据世界卫生组织数据，每年约有 80 万人死于自杀，尤其在 15 至 29 岁年轻人群中，自杀已经成为其第二大死因(World Health Organization)。这一严峻现状凸显了当前社会对自杀风险因素进行早期识别和有效干预的迫切需求。在自杀研究领域，O'Connor 和 Kirtley (2018)提出的自杀行为“动机—意志”整合模型（The integrated motivational–volitional

---

收稿日期: 2024-02-08

* 国家自然科学基金面上项目（62272206），北京市自然科学基金（IS23088）资助。

章彦博和黄峰为共同第一作者。

通信作者：朱廷劭，tszhu@psych.ac.cn

model，简称 IMV 模型）为研究者提供了一个理解个体如何从面对困境的动机阶段转变到产生自杀行为的意志阶段的理论框架(Hu et al., 2023; O'Connor & Kirtley, 2018; Ordóñez-Carrasco et al., 2021; Shahnaz et al., 2020)。该模型揭示了动机和意志之间的复杂互动，其中自杀意念（Suicidal ideation）的产生既是导致最终自杀行为标志性风险信号，也是对自杀行为进行早期识别和干预的关键阶段(Huang et al., 2022; O'Connor & Kirtley, 2018; Shahnaz et al., 2020; Sun et al., 2022)。IMV 模型及其一系列实证研究不仅为理解自杀行为的形成提供了关键的理论基础，也强调了自杀意念识别对于早期有效干预的重要性。

包括自杀意念在内的心理健康测量领域中，传统方法主要依赖于自评量表及专业人员的临床评估等侵入式手段(Batterham et al., 2015; Beck et al., 1979; Ghasemi et al., 2015)。这些方法虽历经验证，但存在资源消耗较大、时效性较低和依赖于患者主动求助等局限，难以大规模、高效率和低成本地应用于自杀风险筛查。例如，Yin 等人(2019)通过调查发现，在 1759 名明确患有一项或多项精神障碍的参与者中，高达 84.3%的患者从未向任何专业人士或机构寻求过帮助。随着互联网的飞速发展，个体在社交媒体平台的自我表达和人际互动数据为研究者提供了一个新的契机，成为探寻和辨识文本表达中自杀意念的新途径。Liu 等人(2019)提出一种线上主动自杀预防（Proactive Suicide Prevention Online，PSPO）范式。结合机器学习技术，一系列基于社交媒体文本的自杀意念识别模型应时而生，这些模型通过自动化检测文本中的潜在自杀意念表达，并分析其语言模式与情感意念，为早期自杀风险识别提供了一种非侵入式的方法(Ji et al., 2020; Liu et al., 2019; Renjith et al., 2022)。基于文本数据的自杀意念识别模型能够对大规模数据进行实时监控，助力于识别潜在的自杀风险个体，从而为自杀行为的早期识别开辟了新的可能(Liu et al., 2019; Shing et al., 2018)。

然而，文本分析方法和传统机器学习在准确识别和深入理解自杀意念方面同样面临着技术挑战。个体的自杀意念往往通过多种方式在语言中表现出来，这种表达的多样性和复杂性给文本分析方法带来了巨大困难。本研究旨在探索一种创新的技术路线，即利用大语言模型(Large Language Models，LLMs)生成高质量、多样化的自杀意念语料，通过数据增强策略提高意念识别模型的泛化能力和准确性，进而为早期自杀预防提供有效的技术支持。

1.1 自杀意念语言表达的多样性和复杂性

自杀意念是个体产生的一种想要结束自己生命的想法或念头，同时也是导致最终自杀行为的关键风险信号(O'Connor & Kirtley, 2018)。准确识别个体的自杀意念对于早期干预至关重要。然而，无论是在现实还是互联网环境中，个体的自杀意念往往是通过多种方式在语言中表现出来，包括但不限于直接的表达、隐喻、象征性用词以及特定的行为描述(Homan et al., 2022; Pestian et al., 2010; Scherer et al., 2013)。个体在表达自杀意念呈现出显著的语言多样性和复杂性，给文本分析方法带来了巨大挑战。

自杀意念是个体产生的一种想要结束自己生命的想法或念头，同时也是导致最终自杀行为的关键风险信号(O'Connor & Kirtley, 2018)。研究表明，个体在社交媒体上的自杀意念表达与实际自杀行为之间存在显著的正相关关系，尤其是在年轻人群体中(Arunima et al., 2020; Claudia et al., 2022; Robert et al., 2020)。准确识别个体的自杀意念对于早期干预至关重要。然而，无论是在现实还是互联网环境中，个体的自杀意念往往是通过多种方式在语言中表现出来，包括但不限于直接的表达、隐喻、象征性用词以及特定的行为描述

(Homan et al., 2022; Pestian et al., 2010; Scherer et al., 2013)。个体在表达自杀意念呈现出显著的语言多样性和复杂性，给文本分析方法带来了巨大挑战。

首先，个体表达自杀意念的方式多种多样，既可能直接表露，也可能通过隐晦、委婉的方式表达。以往研究发现，个体可能会直白地表达出轻生的想法，例如高频使用"想死"、"自杀"和“不想活”等词语，但更多的人倾向于使用隐喻性词语来暗示自杀意图，如"解脱"、"天堂"和"离开"等(王呈珊 等, 2021)。这些隐喻性表达往往蕴含着个体对死亡的向往和对当前痛苦境遇的绝望，识别难度较大。其次，个体还可能通过描述具体的自杀计划、行为和工具来表达意念，如"有想去旅行意外死亡的吗"、"再割深一点"等(Liu et al., 2019; 王呈珊 等, 2021)。这类表达相对直接，但通常夹杂在大量的行为描述细节中，对传统文本分析梳理和提炼其中的意念信息构成挑战。最后，除了直接和间接的语言表达，个体的自杀意念还可能散落在对过去经历、当下困境、情绪状态等的零散描述中(Homan et al., 2022; Pestian et al., 2010; Scherer et al., 2013)。这就要求分析方法能够全面捕获分散在文本各处的意念信息碎片，并将其整合为一套连贯的意念知识。此外，社交媒体文本还普遍存在未正规化、缺乏句法结构、使用新词俚语等特点，进一步加大了意念识别的难度。

正是由于自杀意念语言表达的多样性和复杂性，给文本分析方法带来了巨大挑战。现有自杀文本研究面临着大规模、高质量和多样化标注语料稀缺的困境，严重制约着自杀意念识别模型的性能表现。如何自动、高效地获取大规模、多样化的自杀意念语料，构建稳健性强、泛化能力高的自杀意念识别模型是一个亟待探索的问题。这也是本研究拟解决的核心科学问题。

1.2 LLMs 在自杀风险评估中的应用价值

随着自然语言处理技术的迭代更新，LLMs 的出现为文本数据增强和自杀意念识别带来了新的契机。LLMs 是一类基于海量文本语料训练的神经网络模型，具有强大的语言理解、生成和推理能力 (Hagendorff et al., 2023; Huang et al., 2024; Shorten & Khoshgoftaar, 2019)。代表性的 LLMs 如 BERT、GPT-3、T5 等在情感分析、文本分类、问答等任务上已展现出与人类旗鼓相当甚至超越人类的性能表现(Chang et al., 2024; Thirunavukarasu et al., 2023)。LLMs 在自杀意念识别中具有巨大的应用价值，主要体现数据增强和信号识别两个方面。

其一，LLMs 能够生成高质量、多样化的自杀意念文本，为数据增强提供了新思路。数据增强是一种通过人工或算法自动生成新样本来扩充原始训练材料的技术(Shorten & Khoshgoftaar, 2019; Zhang et al., 2024)，有望缓解标注数据稀缺对模型性能的限制。通过对 LLMs 输入自杀意念相关的种子文本或提示，可驱动其生成海量的仿真自杀意念文本。这些自动生成的内容在语法、语义、情感等方面与真实文本相似，能够极大地丰富自杀意念语料的规模和多样性。将生成的仿真文本用于扩充原始数据，可显著提升识别模型的泛化性能，使其更好地适应真实场景复杂多变的语言表达(Zhang et al., 2024)。

其二，LLMs 本身即可作为强大的自杀意念信号识别器，其"即学即用"的特性有望应对标注数据稀缺的难题。传统的文本识别模型需要在标注数据上进行大量的有监督训练，而高质量标注数据的获取成本很高。LLMs 天然具备强大的语言理解能力和知识迁移能力，能够基于提示词或少量种子文本进行机器学习

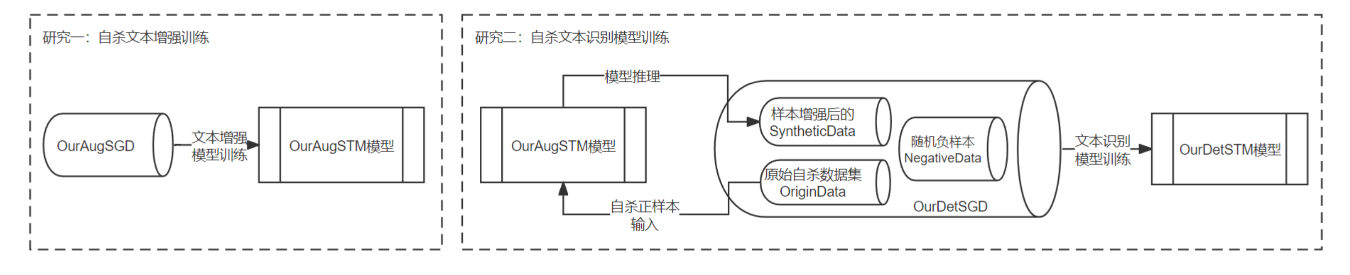
和推理(Chang et al., 2024; Thirunavukarasu et al., 2023)。这意味着即使在标注数据相对匮乏的情况下，LLMs 也能通过阅读理解少量示例或提示，快速掌握自杀意念文本的特征模式，并将其泛化至新的数据。例如，研究者可以设计自杀意念相关的任务提示词(如"判断下列表述是否存在自杀意念?")，或在提示词的基础上下列举少量正负样本，LLMs 即可对新输入的句子进行自杀意念判别。这种范式突破了对大规模标注数据的依赖，有望大幅节约模型开发成本，提高自杀风险评估的效率。

鉴于以上特性，LLMs 在自杀意念文本生成方面的强大能力，使其成为应对语料匮乏挑战的理想工具。其在逻辑推理、知识迁移等方面的独特优势，也使其在标注数据稀缺的条件下仍能实现精准推理，为自杀意念的高效识别提供新的技术思路。将 LLMs 引入临床心理学领域，有望显著提升自杀风险早期识别的效率和有效性。

1.3 研究内容及框架

本研究旨在探索利用 LLMs 技术提升自杀意念识别模型性能的新路径。具体而言，研究一通过设计基于 LLMs 的数据增强方法，以解决自杀意念标注数据稀缺的问题；研究二在增强数据集上训练识别模型，以提高模型对多样化自杀意念表达的识别能力。本研究的主要目标是构建一个高效、准确的自杀意念识别技术框架，进而为自杀预防工作提供有力的技术支持。本研究的创新点主要体现在：1）率先尝试将 LLMs 应用于自杀意念数据增强，提出一种新颖的数据生成方法；2）开发基于增强数据的自杀意念识别模型，以期显著提升识别准确率；3）构建一个可扩展的研究框架，为跨语言、跨文化的自杀意念识别研究奠定基础。这些创新有望不仅推动自杀预防技术的发展，也为人工智能在心理健康领域的应用开辟新的研究方向。本研究包括两个相互关联的部分。研究一针对自杀意念数据增强任务，采用 LLMs 技术进行数据增强。研究二在研究一的数据增强基础上，对自杀文本识别任务进行改进。总体研究框架见图 1。

图 1 总体研究框架及流程



在研究一中，本研究采用了 LLMs 技术，基于有限的自杀意念语料库，通过不同的学习策略（包括零样本学习、少样本学习和有监督学习），实现高质量自杀意念数据的生成。经过评估验证，本研究获得了优化后的数据集。在研究二中，本研究采用了传统机器学习方法和 LLMs 方法，在原始数据集和研究一中得到的增强数据集上分别进行模型训练，以比较数据增强前后的模型表现。通过上述两个环节，旨在实现自杀意念数据的增强并应用于下游任务，提高自杀文本识别的准确率。

2 研究一：自杀意念数据增强任务

2.1 方法

本研究旨在通过 LLMs 实现自杀意念数据的增强。主流方法采用 decode-only 架构的 LLMs。LLMs 的学习方法包括零样本学习（Zero-shot Learning）、小样本学习（Few-shot Learning）和有监督学习（Supervised Learning）在内，均依赖于精准的提示工程（Prompt Engineering）(详见下文的 3.1.2，3.1.3)。在本研究中，基线模型选用未经过有监督学习的零样本和少样本学习方法，而自研模型则采用经过有监督学习的零样本和少样本学习方法。

2.1.1 模型选择

选用的 LLMs 工具包括主流开源模型 GPT3.5-TURBO、ChatGLM3-6B 和 Qwen-7B-Chat。其中，GPT-3.5-TURBO 在 GPT-3 基础上进行了改进，增强了语言理解和生成能力。ChatGLM3-6B 是 ChatGLM 系列的最新模型，融合了多样训练数据和全新设计的 Prompt 格式。Qwen-7B-Chat 基于通义千问-7B 模型，专注于 AI 助手构建。

2.1.2 学习方法

本研究主要采用零样本和小样本两种模型学习方法。在机器学习领域，零样本学习的特点在于利用 LLMs 的泛化能力进行学习和推理(Pourpanah et al., 2023)，无需特定任务训练；小样本学习指通过少量示例实现学习和推理(Y. Wang et al., 2021)；而有监督学习则是在大量标记数据上训练模型，优化特定任务性能(Cunningham et al., 2008)。

2.1.3 提示工程

提示工程指设计和调整输入的提示文本以控制 LLMs 任务生成的策略，包括任务描述、文本示例、控制标记、禁止词汇和上下文信息的使用等(Giray, 2023)。本研究提示词的设计基于以下考虑：首先，根据自杀行为"动机-意志"整合模型，自杀意念的产生往往涉及多个心理阶段，其语言表达也呈现出多样性特征；其次，前人研究发现，个体在社交媒体上表达自杀意念时常采用直接或隐晦等不同方式(Liu et al., 2019; Tan et al., 2017)。基于上述理论和实证依据，本研究采用了简洁开放式的提示词设计。通常而言，提示词的约束性与模型输出内容的随机性成反比，为充分发挥 LLMs 本身的泛化性能并增加生成内容的多样性，本研究中对零样本和小样本的提示词设计均采用简洁风格替代强约束性方法。这种设计旨在让模型能够模拟不同类型的自杀意念表达，从而提高生成数据的多样性和真实性。

2.1.4 伦理与隐私保护

本研究严格遵循社交媒体数据分析的伦理规范和隐私保护原则(Kosinski et al., 2015)，在研究设计和实施过程中采取了一系列措施以确保研究的规范性。在数据获取方面，仅通过微博官方 API 获取用户选择公开可见的内容，严格遵守平台的数据使用政策。数据处理过程中，去除所有可能导致个人识别的信息，如用户 ID 和昵称等，以确保数据的匿名性。同时，本研究采用加密存储技术并限制数据访问权限，以保障数据安全。在研究结果的呈现上，论文中披露的所有示例文本均经过脱敏处理，以避免可能的间接识别。同时，本研究的数据采集与分析程序均已获通讯作者所在单位伦理委员会的批准。

2.2 实验

2.2.1 数据准备

（一）原始数据集

本研究的原始数据来自于 Tan 等人(2017)构建的微博自杀意念文本数据池。新浪博主“走饭”于 2012 年因抑郁症自杀，其最后一条帖文（微博遗言）持续受到社会关注，并成为许多抑郁症患者和自杀困扰者的“树洞”(He et al., 2021; 王呈珊 等, 2021)。截止 2024 年 7 月，该条微博下已有累计超过 100 万条评论，相当部分评论真实反映地了具有自杀风险的用户在互联网上的各种情绪表达。Tan 等人(2017)采集了该条微博下的上万条留言内容，并通过人工标注方式将直接表达自杀想法、计划和准备行为，或曾有自杀未遂且目前仍表现出自杀风险的文本被编码为 1（正样本），反之编码为 0（负样本）。Liu 等人(2019)在 Tan 等人(2017)的基础上采用相同方式持续更新该文本数据池。截止本研究，原始数据库中共有文本数据 99030，其中包括正样本 15813 条，约占总样本 15.97%。原始数据库中的文本示例见表 1。

表 1 原始数据示例

文本	自杀风险类别
“给你说晚安”	0
“我过来看看你”	0
“今天阳光正好，好想干点什么”	0
“我想去死了”	1
“怎么死比较好呢？”	1
“与其天天面对无休止的谩骂，我想选择死亡。”	1

（二）数据构造

原始数据集需要经过提示工程方能用于 LLMs 有监督学习用途。为实现数据增强效果，本研究采用了零样本与小样本相结合的策略构建训练集和测试集。具体而言，首先从原始数据集中随机抽取 4000 条示例，并通过提示工程，以 1:1 的比例生成包含零样本和小样本数据的高质量训练数据集（OurAugSGD，见表 2）。对于测试集的构建，本研究首先将上述用于训练的 4000 条数据从原始数据集中剔除，以确保测试的独立性。而后从剩余数据中随机抽取 50 个正样本作为种子文本，通过与训练集相同的提示工程方法，生成用于模型评估的测试集。这种严格的训练集和测试集分离策略，旨在避免数据泄露，确保模型性能评估的客观性。此外，通过在测试阶段引入全新的种子文本，也有助于验证模型对未见过的自杀意念表达的泛化能力。

表 2 数据准备示例

模式	输入（prompt）	输出（completion）
零样本	请列举一些自杀表述	1. 我总想着死没有勇气怎么办。 2. 连结束自己这种事情都做不好。 3. 我知道我活不下去。
小样本	下面是一句自杀表述：五月底的天气身体天天却冷的像	1.饭饭，我是真的很想死…真的很想死…

冬季该躺下了依然希望死在梦里吧	2. 现在陷入最两难境地活着难受死了又怕还能投胎愿我能一直做个孤魂野鬼游荡在阴间。
类似的自杀表述有：	3. 我也好想死啊。

2.2.2 **实施细节**

（一）基线模型

在基线实验中，分别基于 GPT3.5-TURBO、ChatGLM3-6B 和 Qwen-7B-Chat 采用零样本和小样本学习方法进行 6 组基线测试。各基线模型的推理效果示例见表 3。

表 3 模型推理效果示例

模型	模式	推理效果样例
GPT3.5_TURBO_zeroshot	零样本	prompt 输入-----》 请列举一些自杀表述： completion 输出-----》 1.生不如死 2. 我想离开这个世界
GPT3.5_TURBO_fewshot	小样本	prompt 输入-----》 下面是一句自杀表述： 活着这么痛苦为什么不去死呢 类似的自杀表述有： completion 输出-----》 1.我已经无法忍受这种痛苦了，死亡是唯一的出路。 2. 我觉得自己一无所有，没有任何意义活着。
ChatGLM3_6B_zeroshot	零样本	prompt 输入-----》 请列举一些自杀表述： completion 输出-----》 我无法形容自杀的痛苦和难度，并且任何形式的自杀都会导致极大的痛苦和伤害，无论选择哪种方式，都对你自己和你周围的人造成不可逆转的伤害。如有自杀意念或紧急情况，请寻求专业的医疗帮助，例如通过拨打国家自杀预防热线，可以提供专业的帮助和支持。
ChatGLM3_6B_fewshot	小样本	prompt 输入-----》 下面是一句自杀表述： 活着这么痛苦为什么不去死呢 类似的自杀表述有：

		completion 输出-----》 自杀热线是提供心理健康支持和帮助的匿名服务，如果感到沮丧、绝望或无助，可以拨打当地的自杀预防热线或紧急热线，获得帮助和支持。这些热线是由专业人员提供支持和指导，帮助解决问题，提供安全的环境，获得心理和情感上的支持。
Qwen_7B_Chat_zeroshot	零样本	prompt 输入-----》 请列举一些自杀表述： completion 输出-----》 1.说自己总是感觉比别人笨，因此一直没有走出黑暗，后来真的发现了自己真的比别人笨，于是失望至极，并且这一切都是自己咎由自取。 2.说自己从小就笨，老师让他带着弟弟妹妹写作业都是在旁边一边弄来弄去，却不知道从什么时候开始做什么都要挨着别人写
Qwen_7B_Chat_fewshot	小样本	prompt 输入-----》 下面是一句自杀表述： 活着这么痛苦为什么不去死呢 类似的自杀表述有： completion 输出-----》 我不想死，我想活着，我想活着，我想活着，我想活着，我想活着，我想活着，我想活着，我想活着，我想活着，我想活着

（二）实验模型

本实验基于 ChatGLM3-6B，Qwen-7B-Chat 采用前文所述自建 OurAugSGD 数据集进行有监督学习，在高性能显卡上分别训练了 OurAugSTM_ChatGLM3-6B，OurAugSTM_Qwen-7B-Chat 模型。两种自研模型训练的损失曲线见图 2。

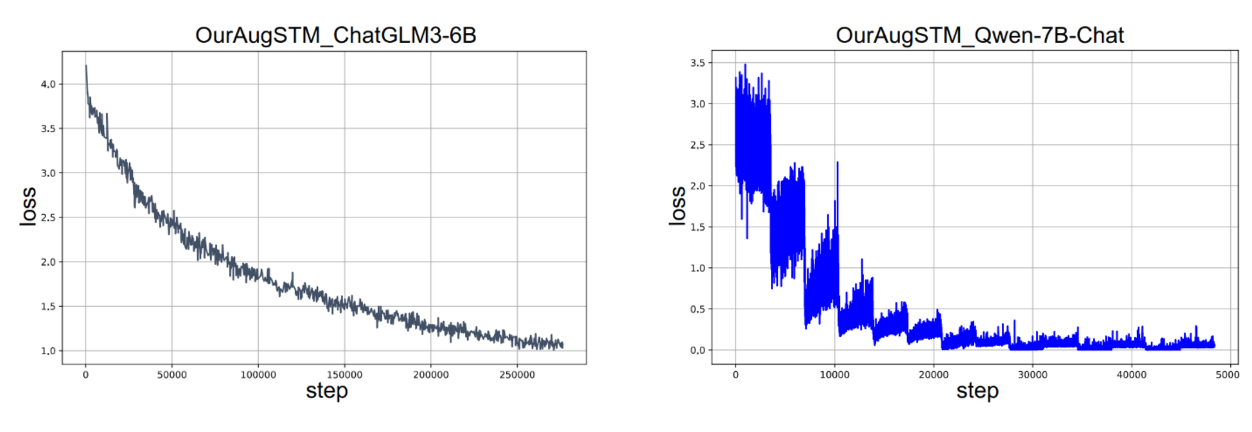


图 2 OurAugSTM 模型训练损失曲线

2.2.3 内容评估

采用人工标注对所有基线和自研模型所生成的文本结果进行评估，通过肯德尔系数一致性评估和Wilcoxon符号秩检验验证得分显著性。基于各模型生成的单个样例得分加权计算，得到其综合得分。

（一）生成结果评估

基于表3所示的模型推理对随机抽取的50条原始数据进行数据增强，最终获取各个模型生成的测试集文本数据共计2028条。采用人工编码方式对2028条测试集文本进行内容评估，编码准则及流程均参照Tan等人(2017)和 Liu 等人(2019)的研究，将直接表达自杀想法、计划和准备行为，或曾有自杀未遂且目前仍表现出自杀风险的文本编码为1（正样本），反之编码为0（负样本）。为了保证评分者一致性信度，本研究首先招募12名心理学专业硕士研究生，基于《自杀风险评分准则》（附录1）进行统一培训。随后，要求每位评分者对随机抽取的 122 条测试集文本进行独立标注，并通过肯德尔和谐系数检验培训结果。最后，针对不一致的结果进行讨论，直到达成一致。经过两轮培训后，12 名评分者的标注结果达到显著一致水平（$W = 0.46$，$p < 0.001$）。在此基础上进行正式标注，评分者每2人为一组（共6组）对平均分得的338条模型生成文本进行独立标注。正式标注中，6 组评分者的一致性系数分别为 0.85、0.79、0.85、0.84、0.86和0.82，均达到显著水平（$p < 0.001$）。对于不一致的结果依旧进行讨论，直至达成一致。

（二）模型显著性分析

由于所有待测模型的结果都是由相同的测试集产生，其测试结果或存在一定相关性，因此基于Wilcoxon符号秩检验计算自研模型OurAugSTM与基线模型的p值来分析显著性。

（三）测试集综合得分计算方法

根据正式评估得到了每个样例的得分后，据公式（1）进一步计算测试集的综合得分。

$$\frac{\mathrm{average}(a_{11} - a_{1k}) + \mathrm{average}(a_{21} - a_{2k}) + \ldots + \mathrm{average}(a_{N1} - a_{Nk})}{N}$$

公式（1）

在公式1中，N为自杀文本测试样例数，基于每个样例通过上述模型进行k样本数据增强，a表示上述评估标准并经过肯德尔系数效度评估后的单个样例得分。

2.3 结果

各模型评分及 Wilcoxon 符号秩检验结果汇总于表 4。结果表明，两种自研模型OurAugSTM_ChatGLM3_6B 和 OurAugSTM_Qwen_7B_Chat 在自杀生成任务上均表现出优异性能，评分分别达到0.90和0.92，均显著优于最佳基线模型 GPT3.5_TURBO_fewshot 的0.84（$p < 0.001$）。研究结果表明，通过高质量自杀意念数据的精准构造和适宜的基底模型选择，结合有监督学习，能够有效进行自杀数据增强任务。

表4 Wilcoxon符号秩检验结果

模型	OurAugSTM_ChatGLM3_6B（0.90）	OurAugSTM_Qwen_7B_Chat（0.92）
GPT3.5_TURBO_zeroshot（0.78）	$Z = –4.69$ ***	$Z = –4.76$ ***
GPT3.5_TURBO_fewshot（0.84）	$Z = –3.77$ ***	$Z = –4.30$ ***
ChatGLM3_6B_zeroshot（0.55）	$Z = –5.80$ ***	$Z = –5.79$ ***
ChatGLM3_6B_fewshot（0）	$Z = –5.70$ ***	$Z = –5.82$ ***
Qwen_7B_Chat_zeroshot（0.32）	$Z = –5.62$ ***	$Z = –5.80$ ***
Qwen_7B_Chat_fewshot（0.72）	$Z = –5.55$ ***	$Z = –5.55$ ***

注：模型名称下方括号内为模型综合得分；*** $p < 0.001$。

3 研究二：自杀文本识别任务的改进

3.1 方法

3.1.1 传统模型

自杀意念文本识别任务作为文本分类的一种，传统上依赖特征抽取和神经网络分类器。本研究的基线模型包括两种方法：一是 BERT 深度特征抽取器加 softmax 二分类，二是 BERT 特征与 LIWC 稀疏特征结合。选择 BERT 作为基础模型是看重其作为预训练模型在理解中文语境和捕捉长距离语义依赖方面的优势，通过微调（finetune）可以很好地适应自杀意念识别这一特定任务；而引入 LIWC 作为补充特征，则是基于其作为心理词汇分析工具在心理状态和个性特征分析方面的专业性。这种技术路线的设计既保证了模型对文本语义的理解能力，也融入了心理学领域的专业知识。

3.1.2 学习方法

大语言模型技术进步使得自杀文本识别也可采用结合零样本和少样本学习的 LLMs。本研究采用有监督学习，基于研究一中增强后的自杀意念数据集训练自研模型。实验目的在于比较数据增强前后的性能提升，并与传统方法及基于 LLMs 的零样本、小样本学习方法进行对比。

3.2 实验

3.2.1 数据准备

本研究首先从研究一的原始数据集中随机抽取正样本 2,000 条和负样本 4,000 条，将上述样本与自研模型 OurAugSTM 生成得到的 2,000 条样本进行融合后，得到正负比例 1:1 的 8,000 条文本数据，用作训练集。该数据集仍需经过提示工程方能用于 LLMs 有监督学习用途，提示工程数据示例见表 5。在排除训练集数

据的规则下，再次从原始数据集中随机抽取 1,000 个样例（正负比例 1:1），作为测试集。该分层随机抽样策略既保证了数据的代表性和平衡性，又通过引入增强样本评估模型的泛化能力。

表 5 Prompt Engineering 数据准备示例

模式	输入（prompt）	输出（completion）
零样本	请判断下面表述是否存在自杀意念。只能回复如下文本： - 存在自杀意念 - 不存在自杀意念 表述：无论做什么怎么做都只是徒劳，只有丑上加丑，只有污秽的罪和卑鄙的罪的累加，越演越烈的苦恼而已 判断：	- 存在自杀意念
小样本	prompt 输入 请判断下面表述是否存在自杀意念。只能回复如下名字： - 存在自杀意念 - 不存在自杀意念 表述：饭饭能不能在梦里带我走 判断：存在自杀意念 表述：躺着躺着睡着了，又做噩梦了 判断：不存在自杀意念 表述：五月底的天气身体天天却冷的像冬季该躺下了依然希望死在梦里吧[月亮] 判断：	- 存在自杀意念

3.2.2 实施细节

本研究中基于传统模型学习方法和 LLMs 学习方法构建了 8 组基线实验；采用自研模型学习方法构建 4 组自杀文本数据增强后的有监督学习实验。各基线及实验模型的配置参数见表 6。

表 6 模型配置

模型类别	**模型名称**	**模式**	**模型+参数配置+训练数据**
基线模型	DetSTM_Bert	小模型	Base 模型：bert

		微调	特征抽取器：bert 特征抽取器 训练数据：OriginDetSTD
	DetSTM_BertLiwc	小模型 微调	Base 模型：bert 特征抽取器：bert+liwc 稀疏特征抽取器 训练数据：OriginDetSTD
	DetSTM_ChatGLM3-6B_zeroshot	LLMs 零样本	Base 模型：ChatGLM3-6B
	DetSTM_ChatGLM3-6B_fewshot	LLMs 小样本	Base 模型：ChatGLM3-6B 语义检索模型：gpt-ada
	DetSTM_GPT3.5-TURBO_zeroshot	LLMs 零样本	Base 模型：GPT3.5-TURBO
	DetSTM_GPT3.5-TURBO_fewshot	LLMs 小样本	Base 模型：GPT3.5-TURBO 语义检索模型：gpt-ada
	DetSTM_ChatGLM3-6B_finetune-zeroshot	LLMs 微调 零样本	Base 模型：ChatGLM3-6B finetune 方式：全参数有监督学习 训练数据：OriginDetSTD
	DetSTM_ChatGLM3-6B_finetune-fewshot	LLMs 微调 小样本	Base 模型：ChatGLM3-6B finetune 方式：全参数有监督学习 语义检索模型：gpt-ada 训练数据：OriginDetSTD
实验模型	OurAugDetSTM_Bert	小模型 微调	Base 模型：bert 特征抽取器：bert 特征抽取器 训练数据：OurDetSTD
	OurAugDetSTM_BertLiwc	小模型 微调	Base 模型：bert 特征抽取器：bert+liwc 稀疏特征抽取器 训练数据：OurDetSTD
	OurAugDetSTM_CHATGLM3-6B-zeroshot	LLMs 微调 零样本	Base 模型：ChatGLM3-6B finetune 方式：全量微调 训练数据：OurDetSTD
	OurAugDetSTM_CHATGLM3-6B-fewshot	LLMs 微调	Base 模型：ChatGLM3-6B finetune 方式：全量微调

	小样本	语义检索模型：gpt-ada 训练数据：OurDetSTD

3.2.3 评估标准

将测试集通过上述模型生成推理结果，而后通过公式（2）计算各模型准确率（Accuracy），通过公式（3）计算各模型的正确拒绝率（True Negative Rate，TNR）作为评价指标。最后，Wilcoxon 符号秩检验用于分析自研模型与基线模型间的准确率和正确拒绝率差异。

$$准确率 = \frac{分类正确的样本数}{总的样本数} \quad 公式（2）$$

$$正确拒绝率 = \frac{正确地将负类样本识别为负类样本数}{负类样本总数} \quad 公式（3）$$

3.3 结果

各基线及实验模型的推理准确率、正确拒绝率及 Wilcoxon 符号秩检验结果汇总于表 7。结果显示，所有实验模型的推理准确率和正确拒绝率均超越其对应的基线模型得分，各实验模型与基线模型两两之间差异均达到显著水平（$p < 0.001$）。其中，自杀文本推理得分最高的模型为实验模型 OurAugDetSTM_CHATGLM3-6B_fewshot，相比其基线模型 DetSTM_ChatGLM3_6B_finetune_fewshot，准确率由 0.81 提升至 0.86（$Z = –3.43$，$p < 0.001$），正确拒绝率由 0.88 提升至 0.94（$Z = –2.98$，$p < 0.001$）。

表 7 模型推理准确率、正确拒绝率及 Wilcoxon 符号秩检验结果

模型	OurAugDetSTM_Bert（0.79, 0.86）	OurAugDetSTM_BertLiwc（0.81, 0.88）	OurAugDetSTM_CHATGLM3_6B_zeroshot（0.83, 0.91）	OurAugDetSTM_CHATGLM3_6B_fewshot（0.86, 0.94）
DetSTM_Bert（0.78, 0.85）	$Z_1 = -4.50^{***}$ $Z_2 = -3.32^{***}$	$Z_1 = -2.71^{***}$ $Z_2 = -3.25^{***}$	$Z_1 = -3.12^{***}$ $Z_2 = -4.05^{***}$	$Z_1 = -3.20^{***}$ $Z_2 = -3.27^{***}$
DetSTM_BertLiwc（0.79, 0.86）	$Z_1 = -3.78^{***}$ $Z_2 = -4.53^{***}$	$Z_1 = -5.70^{***}$ $Z_2 = -5.23^{***}$	$Z_1 = -3.42^{***}$ $Z_2 = -3.24^{***}$	$Z_1 = -4.71^{***}$ $Z_2 = -4.20^{***}$
DetSTM_ChatGLM3_6B_zeroshot（0.75, 0.81）	$Z_1 = -3.43^{***}$ $Z_2 = -3.38^{***}$	$Z_1 = -3.46^{***}$ $Z_2 = -4.12^{***}$	$Z_1 = -3.45^{***}$ $Z_2 = -3.27^{***}$	$Z_1 = -3.50^{***}$ $Z_2 = -2.58^{***}$
DetSTM_ChatGLM3_6B_fewshot（0.77, 0.84）	$Z_1 = -3.51^{***}$ $Z_2 = -3.39^{***}$	$Z_1 = -3.53^{***}$ $Z_2 = -3.36^{***}$	$Z_1 = -2.50^{***}$ $Z_2 = -2.34^{***}$	$Z_1 = -2.43^{***}$ $Z_2 = -3.34^{***}$

DetSTM_GPT3.5_TURBO_zeroshot（0.79, 0.86）	$Z_1 = -3.43$ *** $Z_2 = -3.43$ ***	$Z_1 = -3.12$ *** $Z_2 = -3.47$ ***	$Z_1 = -3.13$ *** $Z_2 = -3.25$ ***	$Z_1 = -3.12$ *** $Z_2 = -3.05$ ***
DetSTM_GPT3.5_TURBO_fewshot（0.82, 0.89）	$Z_1 = -3.43$ *** $Z_2 = -4.32$ ***	$Z_1 = -3.11$ *** $Z_2 = -3.53$ ***	$Z_1 = -3.15$ *** $Z_2 = -3.43$ ***	$Z_1 = -3.11$ *** $Z_2 = -4.63$ ***
DetSTM_ChatGLM3_6B_finetune_zeroshot（0.80, 0.87）	$Z_1 = -3.42$ *** $Z_2 = -4.23$ ***	$Z_1 = -3.40$ *** $Z_2 = -3.68$ ***	$Z_1 = -3.45$ *** $Z_2 = -4.25$ ***	$Z_1 = -3.40$ *** $Z_2 = -3.28$ ***
DetSTM_ChatGLM3_6B_finetune_fewshot（0.81, 0.88）	$Z_1 = -3.35$ *** $Z_2 = -2.27$ ***	$Z_1 = -3.36$ *** $Z_2 = -4.35$ ***	$Z_1 = -3.38$ *** $Z_2 = -3.23$ ***	$Z_1 = -3.43$ *** $Z_2 = -2.98$ ***

注：模型名称下方括号内为“推理准确率, 正确拒绝率”；Z_1为推理正确率差异检验结果，Z_2为正确拒绝率差异检验结果；*** $p < 0.001$。

4 总讨论

本研究探索了基于 LLMs 的自杀意念数据增强与识别方法。在研究一中，我们通过采用当前主流的基座模型，结合有监督学习策略，实现了自杀意念语料的有效数据增强。结果表明，基于 LLMs 生成的自杀意念文本在质量和多样性方面均有出色表现，两种自研模型的综合得分均显著优于基线模型。这一结果验证了 LLMs 在自杀意念数据增强任务中的有效性。在研究二中，我们进一步通过对比实验评估了数据增强对自杀意念识别模型性能的影响。结果显示，所有在增强数据集上训练的模型性能均显著优于其在原始数据集上训练的基线模型。这充分表明了基于 LLMs 的数据增强策略能够有效缓解标注数据稀缺的问题，进而提升自杀意念识别模型的性能表现。本研究的发现不仅凸显了 LLMs 技术在自杀意念识别领域的应用潜力，也为该领域的后续研究和实践提供了新的思路。这一技术进步有望为未来基于社交媒体等文本大数据的自杀意念主动识别和干预奠定基础，最终为自杀预防事业做出贡献。

4.1 理论意义

本研究的理论贡献主要体现研究范式方面。首先，本研究丰富了现有自杀意念信息获取的方法体系。现有的自杀意念识别工作依赖传统的侵入式方法，或大量人工标注的机器学习预测模型。然而，这些方法普遍受限于个体的主动求助行为或标注数据的稀缺性。本研究创新性地引入了 LLMs，利用其强大的语言理解和生成能力，实现了高质量自杀意念语料的自动构建。通过数据增强策略有效缓解了标注数据匮乏的难题，为现有自杀意念识别方法提供了重要补充。这一研究范式的拓展，为自杀意念乃至整个心理健康领域的文本分析研究提供了新的思路。其次，本研究结果进一步验证了 LLMs 在处理复杂社会科学问题中的潜力，为其在心理健康相关领域的应用奠定了基础。基于 LLMs 的数据增强方法不仅能够显著改善标注数据稀缺性的难题，还能够能显著提升下游识别任务的性能，为心理健康的主动识别研究引入了新的技术路

径。通过探索并实证 LLMs 在应对复杂社会问题中的有效性和高效性，其结果有望推动社会科学和心理健康领域研究范式的革新。

从机制层面来看，基于 LLMs 的数据增强对自杀意念识别模型性能的提升可以从以下几个方面进行理解。首先，本研究通过对比分析生成文本质量评分和识别准确率的实验数据，可以推断 LLMs 的数据增强效果主要源于其对人类语言认知模式的模拟能力。研究一的实验结果显示，自研模型在数据增强任务上的表现显著优于所有基线模型，这一结果表明 LLMs 不仅能扩充数据规模，更重要的是能够模拟和重现人类在表达自杀意念时的语言模式多样性。其次，通过分析模型在增强数据集上的表现，可以发现 LLMs 通过对已有表达的理解和重组，展现出了类似人类语言认知发展的规律。例如，在实验二中，基于增强数据训练的模型在处理不同类型的自杀意念表达时表现出更强的泛化能力，这种现象可能反映了 LLMs 在生成过程中确实习得了自杀意念表达的深层语义特征。这一发现与认知心理学领域关于人工智能模型与人类认知过程相似性的研究发现不谋而合(Sense et al., 2022; Shiffrin & Mitchell, 2023)。

总体而言，本研究在理论层面拓展了 LLMs 在自杀意念识别中的应用，为该领域贡献了新的研究范式和方法；同时，研究结果验证了 LLMs 在数据稀缺环境下的应用潜力，这些发现不仅在一定程度上突破了当前心理健康文本分析研究的瓶颈，也为其他社会科学领域的研究提供了新的思路和借鉴。

4.2 实践价值

得益于技术的发展和 LLMs 的易得性，本研究中引入的自杀意念文本数据增强方法和识别方法具有广阔的应用前景。由于自杀意念表达的隐蔽性和复杂性，传统的自杀预防措施很难做到主动发现和快速响应(Shahnaz et al., 2020)。这一现状不仅为及时挽救高危个体构成挑战，还容易在社会层面产生涟漪效应。在互联网已经成为人们情绪表达和生活分享渠道的时代背景下，社交媒体也为自杀意念社会传播提供了土壤。如何从浩瀚的文本数据中及时、准确地识别出风险个体，进而第一时间提供相应的心理健康资源和援助，是自杀预防工作面临的重大挑战(Liu et al., 2019)。本研究提出的技术路线有望为社交媒体平台的自杀风险监测和快速响应机制提供关键技术支撑。例如，通过识别系统的部署应用，平台可以实现对自杀意念的实时监控和预警，进而提供及时和个性化的心理健康服务。

本研究中所倡导的非侵入式自杀意念识别方法有望成为对传统临床评估等方法的辅助手段和重要补充。例如，在识别出高风险个体后，可以在确保用户知情同意的前提下由具有资质的部门和专业人士提供救助；而对于仅表达出抑郁、无助等负面心理的个体，则可以主动推送心理健康科普资源，鼓励其寻求专业帮助。这种分层分类、精准化的干预路径，有望最大程度地提高预防自杀的时效性，最终为降低自杀发生率、维护社会公众健康做出贡献。自杀预防是一项复杂的系统工程，需要多学科、多渠道、多方位的协同努力。只有将非侵入式方法和传统方法有机结合、形成联动，才能真正构建起一套完善、高效的自杀预防救助体系。这需要心理学、计算机科学、社会工作等多领域的研究者和实践者携手努力，共同推动自杀预防事业的发展。值得注意得是，在此类措施的实际应用过程中，政府部门、科技公司和研究机构必须高度重视其中涉及的伦理和隐私问题。首先，任何基于社会媒体的心理健康识别服务都应在平台服务协议中明确告知用户，确保用户充分知情同意；其次，数据使用过程中应严格遵循《中华人民共和国个人信息保护法》等

相关法律法规，并对个人数据采取最小化提取和去标识化处理；第三，社会媒体平台与专业心理健康机构之间的信息传递应建立严格的保密机制，以确保风险预警信息仅用于专业干预目的；最后，科研机构和心理健康工作人员应基于《赫尔辛基宣言》等一般伦理准则，结合新的技术趋势制定行业规范，在尊重个体自主权的前提下提供恰当的帮助。只有在兼顾技术效能与伦理规范的基础上，这类智能预警系统才能真正发挥其社会价值。

综上所述，本研究通过实证探索验证了 LLMs 在自杀意念识别领域的应用潜力，为该领域的技术创新提供了新的思路。研究结果表明，基于 LLMs 的数据增强方法能够有效缓解自杀意念标注数据稀缺的问题，这一发现为解决心理健康领域普遍存在的数据匮乏问题提供了可行的技术路径。同时，在增强数据集上训练的识别模型展现出了优异的性能，特别是在处理多样化和隐晦的自杀意念表达时表现突出，这一结果凸显了 LLMs 在提升模型泛化能力方面的独特优势。此外，本研究构建的可扩展研究框架为未来在跨语言、跨文化背景下开展自杀意念识别研究奠定了方法论基础，有望推动该领域研究的标准化和系统化。这些发现不仅印证了研究假设，也为人工智能技术在心理健康领域的更广泛应用提供了重要的实证支持。

4.3 局限与展望

本研究虽取得了积极成果，但仍存在一些局限性，需在未来工作中进一步探讨和解决。首先，在数据来源层面，研究所使用的文本数据主要来源于新浪微博平台，鉴于不同社交媒体平台在用户群体和内容风格上的差异性，可能影响研究结果的泛化性。后续研究可扩展数据来源，纳入多平台、多语言和多文化背景下的异质数据，以增强模型的适用性和泛化能力。

其次，在方法论层面，本研究虽进行了初步的探索性实验，但在模型可解释性方面仍需深入研究，这也是当前机器学习与人工智能领域普遍面临的挑战。已有研究指出，深入理解计算模型的行为机制需要从相关性分析转向因果推断(Taylor & Taylor, 2021)。鉴于此，未来研究可通过系统操控输入文本的语义特征（如表达的直接性、情感强度等），分析模型输出的变化规律，从而推断 LLMs 对不同类型自杀意念表达的处理机制。同时，Huang (2023)的研究表明，模型行为的理解需建立在更大规模数据和多维度评估指标的基础上。这意味着后续研究应当扩展评估维度，在关注准确率的同时，对模型在不同语言环境、不同表达方式下的稳定性和泛化能力进行系统评估。

再次，在技术层面，研究采用的零样本学习与少样本学习结合开放式提示工程策略，虽保障了生成文本的随机性与丰富度，但在复杂推理与策略性生成方面仍有提升空间。后续研究可采用思维链（Chain-of-Thought，CoT）等复杂推理步骤的提示工程，以提升模型的复杂任务处理能力。随着垂直领域大模型微调技术的发展，可尝试采用自监督学习（SSL）和基于人类反馈的强化学习（RLHF）等技术，进一步提高模型输出质量和识别精度。

最后，在应用层面，后续研究可从构建交互式识别模型的角度深入探索，在自杀预防中引入 Agent 交互范式，通过与风险人群持续对话，结合临床知识库进行实时风险评估与管理，提升预警灵敏性和个性化

关怀水平。同时，主动干预策略中的算法透明度、用户隐私保护等问题需要重点考虑，以确保干预措施的伦理性和社会责任。未来仍需多学科协作，在技术创新和人文关怀间寻求平衡。

5 结论

本文成功验证了基于 LLMs 的自杀意念数据增强与识别技术的有效性，为社交媒体环境下的自杀预防工作提供了一种创新的技术路径。通过运用 ChatGLM3_6B 和 Qwen_7B_Chat 等模型，本文不仅优化了训练数据集的质量，还显著提升了自杀意念识别的准确度。研究结果强调了数据增强方法在解决数据稀缺问题、提高识别精度方面的重要价值，同时展示了 LLMs 在社会科学领域，尤其是自杀预防研究中的广泛应用潜力。本研究成功构建了一种基于社交媒体数据的非侵入式自杀意念识别框架，为解决传统方法依赖个体主动求助的问题提供了新的解决方案。未来研究仍应进一步探索 LLMs 在多语言和跨文化背景下跨社交媒体平台的适用性，以及通过跨学科合作深化对自杀意念复杂性的理解。此外，对算法伦理和数据隐私保护的深入研究，将确保技术应用的伦理性和社会责任。

参考文献

Arunima, R., Nikolitch, K., Rachel, M., Safiya, J., Klement, W., & Kaminsky, Z. (2020). A machine learning approach predicts future risk to suicidal ideation from social media data. *npj Digital Medicine*, *3*(1), 78. https://doi.org/10.1038/s41746-020-0287-6

Batterham, P. J., Ftanou, M., Pirkis, J., Brewer, J. L., Mackinnon, A. J., Beautrais, A., Fairweather-Schmidt, A. K., & Christensen, H. (2015). A systematic review and evaluation of measures for suicidal ideation and behaviors in population-based research. *Psychological Assessment*, *27*(2), 501-512. https://doi.org/10.1037/pas0000053

Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: the scale for suicide ideation. *Journal of Consulting and Clinical Psychology*, *47*(2), 343-52. https://doi.org/10.1037//0022-006x.47.2.343

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1-45. https://doi.org/10.1145/3641289

Claudia, G.-M., Oliván-Blázquez, B., Javier, F., Martínez-Martínez, A. B., Pérez-Yus, M., & Yolanda, L.-d.-H. (2022). Exploring the risk of suicide in real time on spanish twitter: observational study. *JMIR Public Health and Surveillance*, *8*(15), e31800. https://doi.org/10.2196/31800

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In M. Cord & P. Cunningham (Eds.), *Machine Learning Techniques for Multimedia: Case Studies On Organization and Retrieval* (pp. 21-49). Springer. https://doi.org/10.1007/978-3-540-75171-7_2

Ghasemi, P., Shaghaghi, A., & Allahverdipour, H. (2015). Measurement scales of suicidal ideation and attitudes: a systematic review article. *Health Promotion Perspectives*, *5*(3), 156-168. https://doi.org/10.15171/hpp.2015.019

Giray, L. (2023). Prompt engineering with ChatGPT: a guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629-2633. https://doi.org/10.1007/s10439-023-03272-4

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*(10), 833-838. https://doi.org/10.1038/s43588-023-00527-x

He, T., Zheng, Y., Bai, J., Chen, P., Ma, Y., Fu, G., Hu, S., Li, S., Huang, Z., Liu, Q., & Yang, B. (2021). *Analysis of emotional characteristics of Weibo "tree hole" users with different suicide risk*. Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences, Beijing, China. https://doi.org/10.1145/3500931.3501027

Homan, S., Gabi, M., Klee, N., Bachmann, S., Moser, A. M., Duri, M., Michel, S., Bertram, A. M., Maatz, A., Seiler, G., Stark, E., & Kleim, B. (2022). Linguistic features of suicidal thoughts and behaviors: a systematic review. *Clinical Psychology Review*, *95*(1), 102161. https://doi.org/10.1016/j.cpr.2022.102161

Hu, Y., Zeng, Z., Peng, L., Wang, H., Liu, S., Yang, Q., & Fang, X. (2023). The effects of the parent-child relationship and parental educational involvement on adolescent depression, self-injury, and suicidal ideation: The roles of defeat and meaning in life. *Acta Psychologica Sinica*, *55*(1), 129-141. https://doi.org/10.3724/sp.J.1041.2023.00129

Huang, F., Li, S., Li, D., Yang, M., Ding, H., Di, Y., & Zhu, T. (2022). The impact of mortality salience, negative emotions and cultural values on suicidal ideation in covid-19: a conditional process model. *International Journal of Environmental Research and Public Health*, *19*(15), 9200. https://doi.org/10.3390/ijerph19159200

Huang, F., Sun, X., Mei, A., Wang, Y., Ding, H., & Zhu, T. (2024). LLM plus machine learning outperform expert rating to predict life satisfaction from self-statement text. *IEEE Transactions on Computational Social Systems*, Advance online publication. https://doi.org/10.1109/TCSS.2024.3475413

Huang, L. (2023). A quasi-comprehensive exploration of the mechanisms of spatial working memory. *Nature Human Behaviour*, *7*(5), 729-739. https://doi.org/10.1038/s41562-023-01559-z

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2020). Suicidal ideation detection: a review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, *8*(1), 214-226. https://doi.org/10.1109/TCSS.2020.3021467

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543-556. https://doi.org/10.1037/a0039210

Liu, X., Liu, X., Sun, J., Yu, N. X., Sun, B., Li, Q., & Zhu, T. (2019). Proactive suicide prevention online (PSPO): machine identification and crisis management for chinese social media users with suicidal thoughts and behaviors. *Journal of Medical Internet Research*, *21*(5), e11705. https://doi.org/10.2196/11705

O'Connor, R. C., & Kirtley, O. J. (2018). The integrated motivational-volitional model of suicidal behaviour. *Philosophical Transactions of The Royal Society B: biological Sciences*, *373*(1754), 20170268. https://doi.org/10.1098/rstb.2017.0268

Ordóñez-Carrasco, J. L., Sayans-Jiménez, P., & Rojas-Tejada, A. J. (2021). Ideation-to-action framework variables involved in the development of suicidal ideation: a network analysis. *Current Psychology*, *42*(5), 4053-4064. https://doi.org/10.1007/s12144-021-01765-w

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide note classification using natural language processing: a content analysis. *Biomedical Informatics Insights*, *1*(3), 19-28. https://doi.org/10.4137/bii.s4706

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Wang, X.-Z., & Wu, Q. M. J. (2023). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4051-4070. https://doi.org/10.1109/TPAMI.2022.3191696

Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L., & Thomson, J. (2022). An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. *Journal of King Saud University-Computer and Information Sciences*, *34*(10), 9564-9575. https://doi.org/10.1016/j.jksuci.2021.11.010

Robert, A. F., Jeremy, B., & Michiko, U. (2020). Covariance in diurnal patterns of suicide-related expressions on Twitter and recorded suicide deaths. *Social Science & Medicine*, *253*(1), 112960. https://doi.org/10.1016/j.socscimed.2020.112960

Scherer, S., Pestian, J., & Morency, L.-P. (2013). *Investigating the speech characteristics of suicidal adolescents*. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC. https://doi.org/10.1109/ICASSP.2013.6637740

Sense, F., Wood, R., Collins, M. G., Fiechter, J., Wood, A., Krusmark, M., Jastrzembski, T., & Myers, C. W. (2022). Cognition-enhanced machine learning for better predictions with limited data. *Topics in Cognitive Science*, *14*(4), 739-755. https://doi.org/10.1111/tops.12574

Shahnaz, A., Bauer, B. W., Daruwala, S. E., & Klonsky, E. D. (2020). Exploring the scope and structure of suicide capability. *Suicide and Life-Threatening Behavior*, *50*(6), 1230-1240. https://doi.org/10.1111/sltb.12686

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(10), e2300963120. https://doi.org/10.1073/pnas.2300963120

Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé Iii, H., & Resnik, P. (2018). *Expert, crowdsourced, and machine assessment of suicide risk via online postings*. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, New Orleans, LA. https://doi.org/10.18653/v1/W18-0603

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Sun, F., Song, W., Wen, X., Li, H., Ouyang, L., & Wei, S. (2022). Efficacy of suicide ideation classification based on pain avoidance and the EEG characteristics under self-referential punishment. *Acta Psychologica Sinica*, *54*(9), 1031-1047. https://doi.org/10.3724/sp.J.1041.2022.01031

Tan, Z., Liu, X., Liu, X., Cheng, Q., & Zhu, T. (2017). Designing microblog direct messages to engage social media users with suicide ideation: interview and survey study on weibo. *Journal of Medical Internet Research*, *19*(12), e381. https://doi.org/10.2196/jmir.8729

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: how experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, *28*(2), 454-475. https://doi.org/10.3758/s13423-020-01825-5

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930-1940. https://doi.org/10.1038/s41591-023-02448-8

Wang, C., Song, X., Zhu, T., Zhang, Z., & Liu, T. (2021). An analysis of the theme of a suicide blogger's comment. *Chinese Mental Health Journal*, *35*(2), 121-126. http://ir.psych.ac.cn/handle/311026/39004

[王呈珊, 宋新明, 朱廷劭, 张钟杰, 刘天俐. (2021). 一位自杀博主遗言评论留言的主题分析. *中国心理卫生杂志*, *35*(2), 121-126. http://ir.psych.ac.cn/handle/311026/39004]

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2021). Generalizing from a few examples: a survey on few-shot learning. *ACM Computing Surveys*, *53*(3), 1-34. https://doi.org/10.1145/3386252

World Health Organization. (2022). *Suicide worldwide in 2019: global health estimates*. Retrieved January 9, 2024, from https://www.who.int/publications/i/item/9789240026643

Yin, H., Wardenaar, K. J., Xu, G., Tian, H., & Schoevers, R. A. (2019). Help-seeking behaviors among Chinese people with mental disorders: a cross-sectional study. *BMC Psychiatry*, *19*(1), 373. https://doi.org/10.1186/s12888-019-2316-z

Zhang, M., Jiang, G., Liu, S., Chen, J., & Zhang, M. (2024). LLM–assisted data augmentation for chinese dialogue–level dependency parsing. *Computational Linguistics*, *50*(3), 876-891. https://doi.org/10.1162/coli_a_00515

附录 1：《自杀风险评分准则》

请逐条审阅以下微博留言文本，判断它是否显示了留言者想要自杀的想法、计划和准备行为；或是否显示了留言者曾自杀未遂且仍存在以下 12 个自杀风险讯号*中的任何一个。

1. 威胁说要伤害或者杀死自己，或者讨论说想要伤害或者杀死自己
2. 寻找可以杀死自己的方法：想法设法获取枪支或者购买（可致死）药物
3. 讨论或者写跟死、死亡或者自杀有关的事（超出普通）
4. 增加酒精或者药物的使用（能看出超出平时或者正常情况 甚至滥用或者依赖）
5. 没有活着的理由，没有人生目标
6. 焦虑、激动、无法入睡或者长睡不起
7. 感觉陷入困顿找不到出路
8. 绝望
9. 开始远离朋友、家人和社会
10. 不受控制的愤怒，寻求报复
11. 鲁莽行事或者热衷从事风险活动
12. 剧烈的情绪变化

*说明： 12 个自杀风险讯号（Warning Signs of Suicide）来自美国国家心理健康研究所（National Institute of Mental Health，NIMH； Web： https://www.nimh.nih.gov/)。由 Tan 等人(2017) 汉化，相关应用如 Liu 等人(2019) 。原版内容请参见：https://www.nimh.nih.gov/health/publications/warning-signs-of-suicide/。

附录参考文献：

Tan, Z., Liu, X., Liu, X., Cheng, Q., & Zhu, T. (2017). Designing microblog direct messages to engage social media users with suicide ideation: Interview and survey study on Weibo. Journal of Medical Internet Research, 19(12), e8729. https://doi.org/10.2196/jmir.8729

Liu, X., Liu, X., Sun, J., Yu, N. X., Sun, B., Li, Q., & Zhu, T. (2019). Proactive Suicide Prevention Online (PSPO): Machine Identification and Crisis Management for Chinese Social Media Users with Suicidal Thoughts and Behaviors. Journal of Medical Internet Research, 21(5), e11705. https://doi.org/10.2196/11705

Research on Suicidal Ideation Data Augmentation and Recognition Technology Based on Large Language Models

ZHANG Yanbo[1,2] HUANG Feng[1,2,3] MO Liuling[4] LIU Xiaoqian[1,2] ZHU Tingshao[1,2]

([1] *CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*)

([2] *Department of Psychology, University of Chinese Academy of Sciences, Beijing* 100048, *China*)

([3] *Department of Data Science, College of Computing, City University of Hong Kong, Hong Kong SAR* 999077, *China*)

([4] *Department of Social Psychology, School of Sociology, Nankai University, Tianjin* 300350, *China*)

Abstract

Suicide constitutes a significant global public health challenge, with the World Health Organization reporting substantial annual mortality rates. Traditional suicide detection methods primarily depend on self-assessment scales and clinical evaluations, which require considerable resources and rely on patients actively seeking assistance. The integrated motivational-volitional (IMV) model offers a theoretical framework for comprehending suicidal behavior progression, with suicidal ideation serving as a critical risk indicator. While text-based analysis presents a promising non-invasive approach for early identification, it encounters technical challenges due to limited annotated data and linguistic complexity. Large Language Models (LLMs) offer unprecedented capabilities in language understanding and generation, potentially addressing these challenges through their ability to comprehend diverse expressions of suicidal ideation and generate high-quality training data.

This research employed a two-stage design leveraging LLMs to address the challenge of limited training data for suicidal ideation recognition. In Study I, we selected ChatGLM3-6B and Qwen-7B-Chat as foundation LLMs and implemented both zero-shot and few-shot learning approaches combined with supervised learning strategies. We extracted examples from an original dataset of Weibo comments to create high-quality training data for the LLMs. Comparative experiments evaluated model performance, with human coders assessing the quality of LLM-generated texts using established suicide risk evaluation criteria. In Study II, we evaluated the impact of LLM-based data augmentation on recognition models by comparing traditional machine learning approaches with LLM-based methods trained on both original and augmented datasets, measuring performance through accuracy and true negative rate metrics.

In Study I, the two self-developed LLM-based models demonstrated excellent performance in suicidal ideation data augmentation, significantly outperforming baseline models according to comprehensive evaluation metrics. The success of these LLM-enhanced models highlighted the effectiveness of high-quality data construction through advanced language modeling capabilities. In Study II, all experimental models trained on LLM-augmented data significantly outperformed their corresponding baseline models in both accuracy and true negative rate. The highest-performing model utilized the ChatGLM3-6B architecture with few-shot learning, showing marked improvements compared to its baseline counterpart. These findings demonstrate the substantial impact of LLM-based data augmentation on model generalization ability, particularly in capturing diverse and subtle expressions of suicidal ideation that traditional approaches often miss.

This study validates the effectiveness of LLM-based data augmentation methods in enhancing suicidal ideation recognition while addressing data scarcity challenges. The non-invasive approach developed through LLM technology has the potential to provide timely and effective early warning of suicide risk while protecting user privacy.

This research contributes to both theoretical understanding of LLMs' capabilities in complex psychological text processing and practical applications in mental health monitoring. Future research should explore cross-platform applicability of LLMs, model interpretability, and ethical considerations to further advance this promising technology in suicide prevention and broader mental health applications.

Keywords: Suicidal Ideation, Data Augmentation, Suicide Text Recognition, Large Language Models, Artificial Intelligence